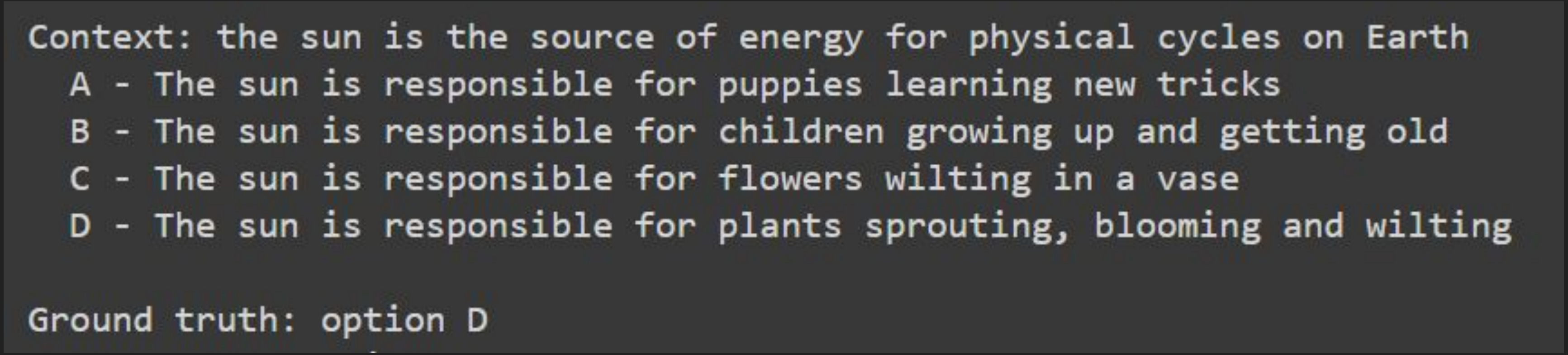
Project: Commonsense Question Answering

Group 2: Manikanta Mandlem, Meetu Patel, Nihaal Subhash, Sai Indraneel Amara

The Dataset - OpenbookQA

- Question-answering dataset modeled after open book exams
- 5,957 multiple-choice elementary-level science questions
- 4,957 train, 500 dev, 500 test

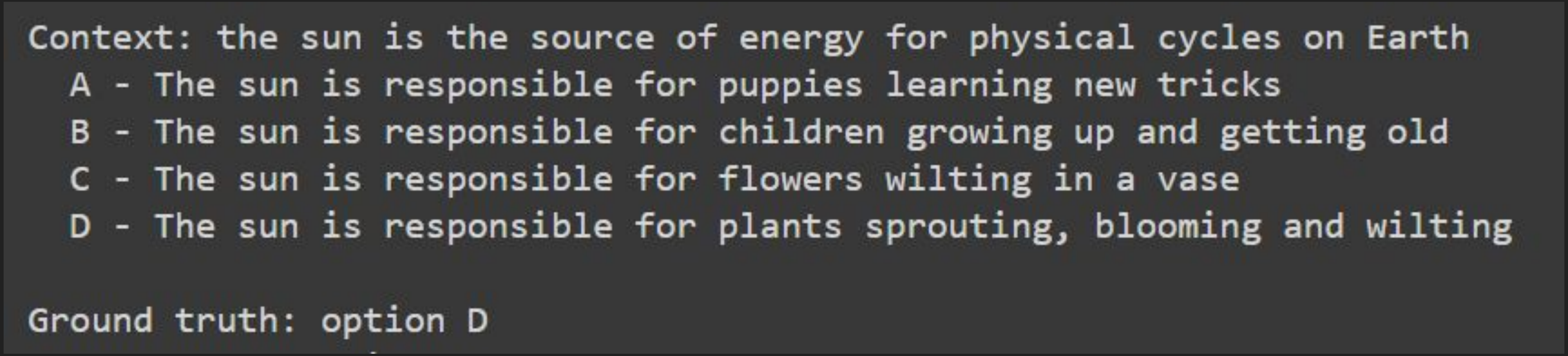


```
Context: the sun is the source of energy for physical cycles on Earth
  A - The sun is responsible for puppies learning new tricks
  B - The sun is responsible for children growing up and getting old
  C - The sun is responsible for flowers wilting in a vase
  D - The sun is responsible for plants sprouting, blooming and wilting

Ground truth: option D
```

The Dataset - OpenbookQA

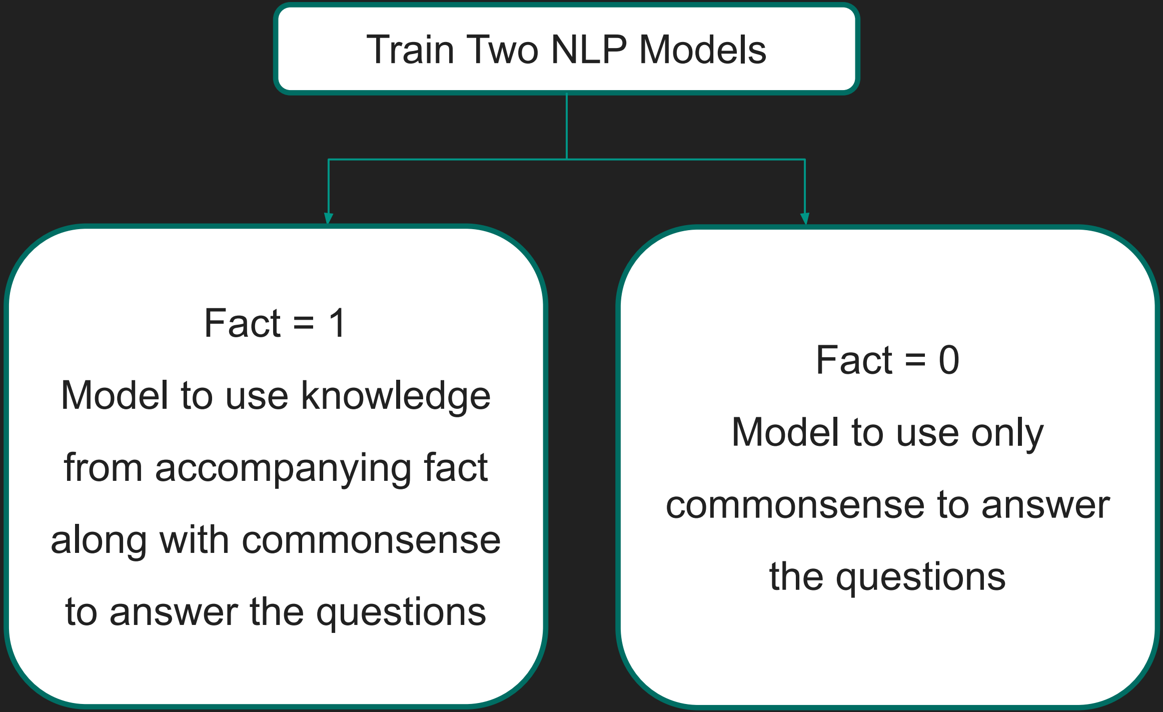
- Probe the understanding of a small “book” of 1,326 core science facts and the application of these facts to novel situations
- External knowledge would help in answering these questions



```
Context: the sun is the source of energy for physical cycles on Earth
  A - The sun is responsible for puppies learning new tricks
  B - The sun is responsible for children growing up and getting old
  C - The sun is responsible for flowers wilting in a vase
  D - The sun is responsible for plants sprouting, blooming and wilting

Ground truth: option D
```

The Task



Train Two NLP Models

Fact = 1

Model to use knowledge from accompanying fact along with commonsense to answer the questions

Fact = 0

Model to use only commonsense to answer the questions

Approaches

Approach 1: Eliminate an Option

using explicit method to eliminate an option and retrain the model with the remaining three options

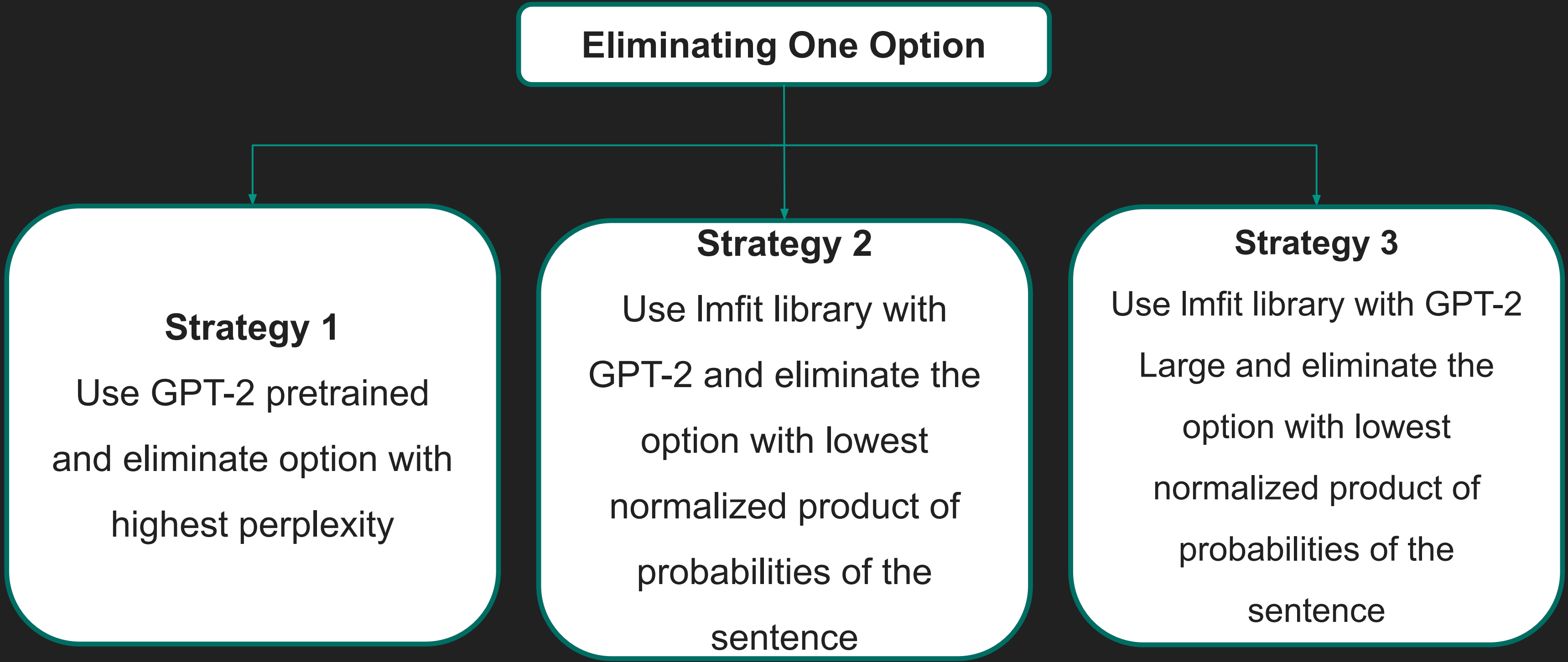
The River Foyle is found in which part of the United Kingdom?
A: ~~England~~
B: Scotland
C: Northern Ireland
D: Wales

Approaches

Approach 2: External Knowledge Base

Pretrain the model on external knowledge and finetune it for OpenBookQA

Approach 1



Eliminating One Option

Strategy 1

Use GPT-2 pretrained and eliminate option with highest perplexity

Strategy 2

Use lmfit library with GPT-2 and eliminate the option with lowest normalized product of probabilities of the sentence

Strategy 3

Use lmfit library with GPT-2 Large and eliminate the option with lowest normalized product of probabilities of the sentence

Approach 1: Results

Strategy	Approach	Probability of retaining correct option
0	Random (Baseline for comparison)	0.75
1	Remove using perplexity	0.73
2	Remove using GPT2 and normalized product of probabilities	0.82
3	Same as (2) but with GPT2 Large	0.80

Approach 1: Results

We selected the ‘Remove using GPT2 product of token probabilities normalized by sentence length’ approach as it had the highest probability (82%) of retaining the correct option. We trained our current model to choose the correct option amongst the three options. The results are as follows:

Sr. No.	Approach	Fact = 0	Fact = 1
1	Baseline	0.550	0.691
2	Explicitly removed 1 option	0.453	0.518

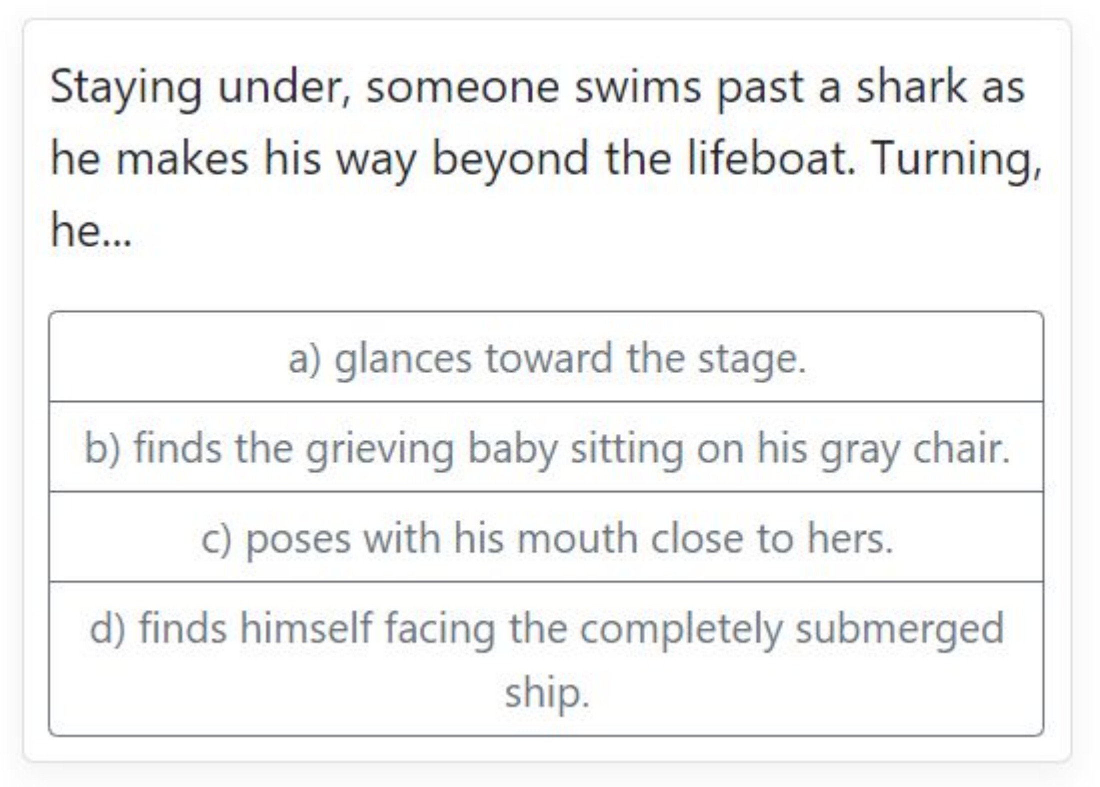
Approach 2:

- We attempted to fine tune using 3 different datasets:
    1. SWAG
    2. CommonSense QA
    3. Allen AI Science Challenge

SWAG (Situations With Adversarial Generations)

- This is a large-scale dataset for the task of grounded commonsense inference
- Example - ”She opened the hood of the car," □ “and examined the engine”
- Consists of 113k multiple choice questions about grounded situations
- Each question is a video caption from LSMDC or ActivityNet Captions, with four answer choices about what might happen next in the scene
- correct answer is the (real) video caption for the next event in the video
- the three incorrect answers are adversarially generated and human verified

SWAG (Situations With Adversarial Generations)



Staying under, someone swims past a shark as he makes his way beyond the lifeboat. Turning, he...

a) glances toward the stage.

b) finds the grieving baby sitting on his gray chair.

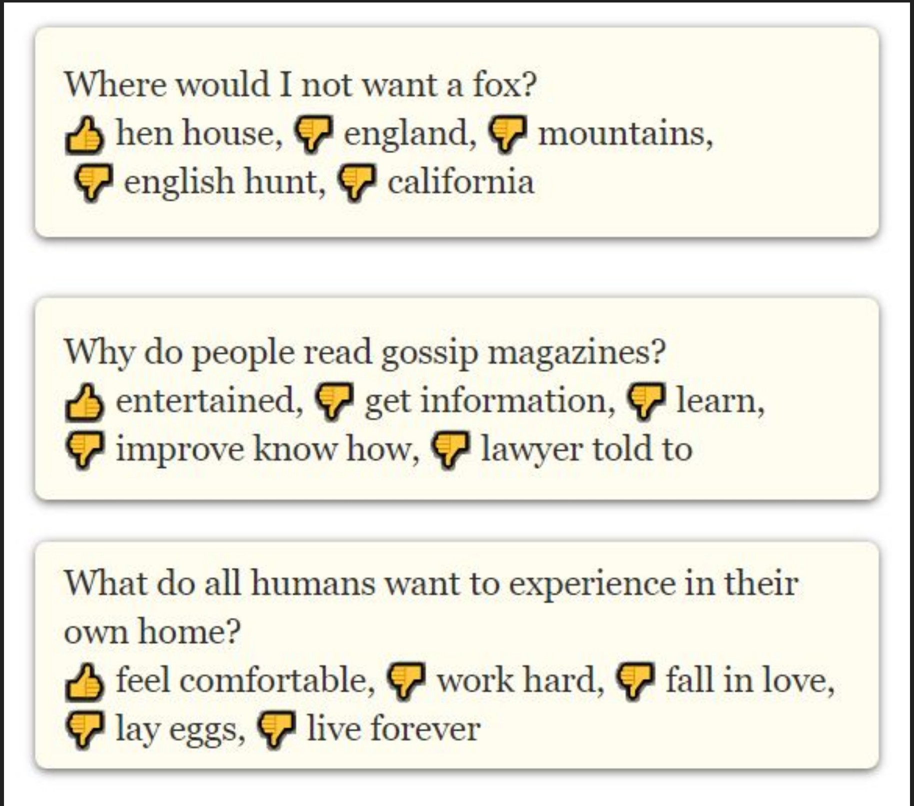
c) poses with his mouth close to hers.

d) finds himself facing the completely submerged ship.

CommonsenseQA

- This is a multiple-choice question answering dataset
- Require commonsense knowledge to answer correctly
- Consists of 12,102 questions with one correct choice and four distractors
- Customized this dataset to our use case by removing one distractor
- Pre-trained using all train, valid and test datasets
- Then incrementally finetuned to OpenbookQA dataset

CommonsenseQA



Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california

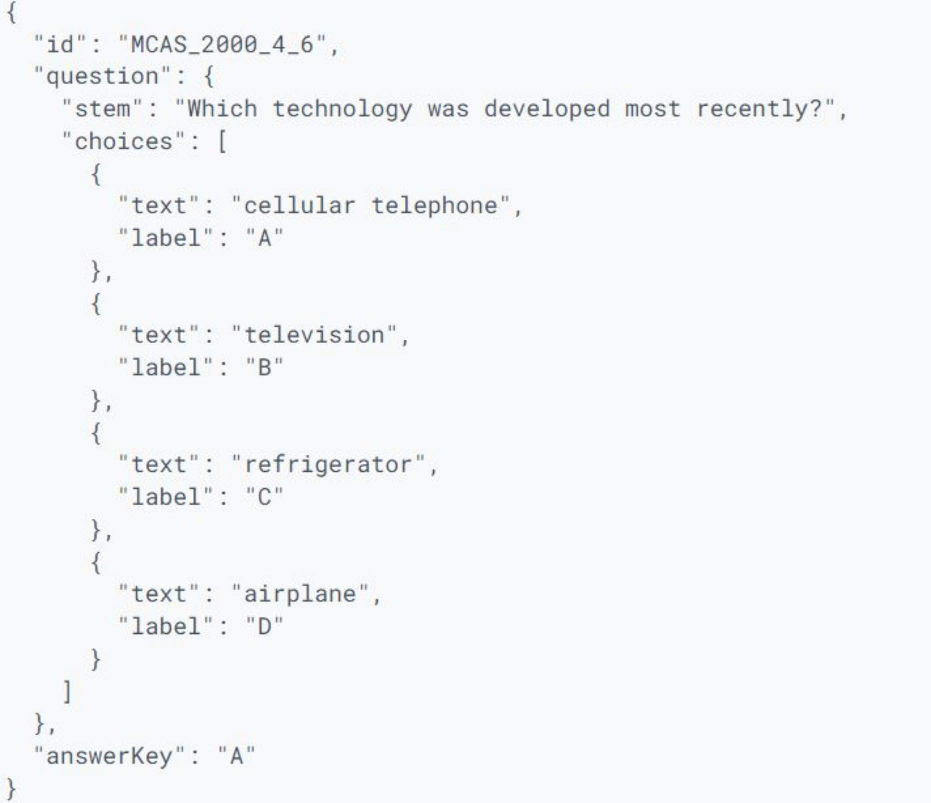
Why do people read gossip magazines?
👍 entertained, 👎 get information, 👎 learn,
👎 improve know how, 👎 lawyer told to

What do all humans want to experience in their own home?
👍 feel comfortable, 👎 work hard, 👎 fall in love,
👎 lay eggs, 👎 live forever

Allen AI Science Challenge

- 7,787 genuine grade-school level, multiple-choice science questions
- Assembled to encourage research in advanced question-answering
- Partitioned into a Challenge Set and an Easy Set
- Idea is that this dataset would help imbibe some general knowledge to the model which would assist the model in common sense reasoning tasks, especially when the relevant fact is not provided

Allen AI Science Challenge



```
{
  "id": "MCAS_2000_4_6",
  "question": {
    "stem": "Which technology was developed most recently?",
    "choices": [
      {
        "text": "cellular telephone",
        "label": "A"
      },
      {
        "text": "television",
        "label": "B"
      },
      {
        "text": "refrigerator",
        "label": "C"
      },
      {
        "text": "airplane",
        "label": "D"
      }
    ]
  },
  "answerKey": "A"
}
```

Approach 2: Results

Approach	**Accuracy for Fact = 0**	**Accuracy for Fact = 1**
Baseline	0.55	0.69
Fine Tuning on Pooled Swag and OpenBookQA	0.56	0.69
Pre-training on Swag and incrementally fine tuning on OpenBookQA	0.56	0.69
Pre-training on CommonsenseQA and incrementally fine tuning on OpenBookQA	**0.58**	**0.71**
Fine Tuning on Allen AI Science Challenge (Easy) and incrementally fine tuning on OpenBookQA	0.55	0.71
Fine Tuning on Pooled CommonsenseQA + Allen AI and incrementally fine tuning on OpenBook QA	0.56	0.70

Areas of Opportunity

- Use the representations generated by GPT-2 (eg: the context embeddings or logits) to learn a classifier to eliminate one incorrect choice.
- Train a model to look at the question and retrieve the relevant fact from the crowdsourced list of facts

SOURCE CODES (NOT EXHAUSTIVE)

- Baseline Code:

  https://colab.research.google.com/drive/1YdsFtrPED89UpFMMFFWIY6mKzasnzLcJ

- Code to produce datasets for pretraining (approach 2):

  https://colab.research.google.com/drive/1NOw7A7wYfYq2ERbK1V8AYzHbRo1UFeVf

- Code to finetune on different datasets then incrementally fine tune:

  https://colab.research.google.com/drive/1P9HDI5b2pJLOYG9EzRO8sxp4BosR1mxL#scrollTo=iUF4AD_wY_-P

THANK YOU

Manikanta Mandlem
Meetu Patel
Nihaal Subhash
Sai Indraneel Amara